

Chapter 10

Alternative Methods for Regression

Concentrations appear linearly related to distance down-dip in an aquifer. OLS regression shows the residuals to be of generally constant variance. However, several outliers in the data set inflate the standard error, and what appears graphically as a strong linear relationship tests as being insignificant due to the outliers' influence. How can a more robust linear fit be obtained which is not overly sensitive to a few outliers, and describes the linear relation between concentration and distance?

A water supply intake is to be located in a stream so that water elevation (stage) is below the intake only 5 percent of the time. Monitoring at the station is relatively recent, so OLS relating this and a nearby site having a 50 year record is used to generate a pseudo 50-year stage record for the intake station. The 5th percentile of the pseudo record is used as the intake elevation. Given that OLS estimates are reduced in variance compared to actual data, this elevation estimate will not be as extreme as it should be. What alternatives to OLS would provide better estimates?

The mass of a radionuclide present within the aquifer of one county was computed by performing a regression of concentration versus log of the hydraulic conductivity measured at 20 wells. This equation was used to generate estimates at 100 locations of known hydraulic conductivity, which are then multiplied by the volumes of water, and summed. However, the regression equation shows a marked increase in variance of concentration with increasing conductivity, even though the relationship is linear. Transformations may produce a nonlinear relationship, with probable transformation bias. An alternative to OLS is therefore required to account for heteroscedasticity without employing a transformation.

Situations such as the above frequently arise where the assumptions of constant variance and normality of residuals required by OLS regression are not satisfied, and transformations to remedy this are either not possible, or not desirable. In addition, the inherent reduction in variance of OLS estimates is not appropriate when extending records. In these situations, alternative methods are better for fitting lines to data. These include nonparametric rank-based methods, lines which minimize other than the squared residuals, and smooths.

10.1 Kendall-Theil Robust Line

The significance of a linear dependence between two continuous variables Y and X or their transforms may be tested by determining whether the regression slope coefficient for the explanatory variable is significantly different from zero. This is equivalent to the test for significance of the linear correlation coefficient r between Y and X . In a similar fashion, Kendall's rank correlation coefficient τ (see Chapter 8) may be used to test for any monotonic, not just linear, dependence of Y on X . Related to τ is a robust nonparametric line applicable when Y is linearly related to X . This line will not depend on the normality of residuals for validity of significance tests, and will not be strongly affected by outliers, in contrast to OLS regression.

The robust estimate of slope for this nonparametric fitted line was first described by Theil (1950). An estimate of intercept is also available (Conover, 1980, p. 267). Together these define an estimate of a complete linear equation of the form:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 \cdot X$$

This line is closely related to Kendall's τ , in that the significance of the test for H_0 : slope $\beta_1=0$ is identical to the test for H_0 : $\tau=0$.

10.1.1 Computation of the Line

The Theil slope estimate \hat{b}_1 is computed by comparing each data pair to all others in a pairwise fashion. A data set of n (X,Y) pairs will result in $n(n-1)/2$ pairwise comparisons. For each of these comparisons a slope $\Delta Y/\Delta X$ is computed (figure 10.1). The median of all possible pairwise slopes is taken as the nonparametric slope estimate \hat{b}_1 .

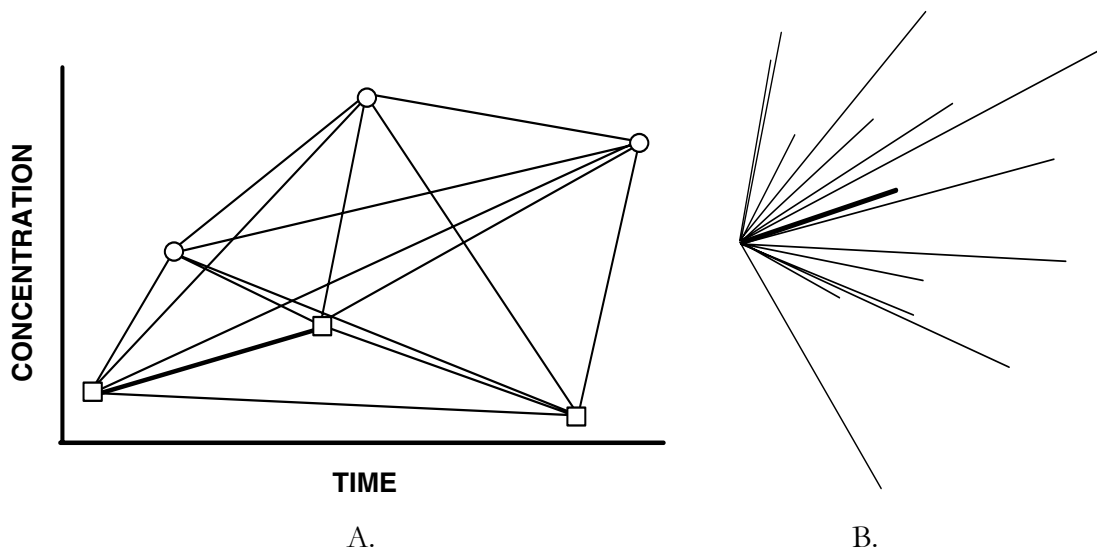


Figure 10.1

- A. All possible pairwise slopes between six data points.
 B. All possible slopes rearranged to meet at a common origin
 The thick line is the median of the 15 slopes.

the data value were changed from 16 to 200, b_1 would be greatly inflated while \hat{b}_1 would again remain at 1. The estimator \hat{b}_1 is clearly resistant to outliers. It responds to the bulk of the data.

\hat{b}_1 is an unbiased estimator of the slope of a linear relationship, and b_1 from OLS is also an unbiased estimator. However, the variance of the estimators differ. When the departures from the true linear relationship (true residuals) are normally distributed, OLS is slightly more efficient (has lower variance) than the Kendall-based line. When residuals depart from normality (are skewed or prone to outliers), then \hat{b}_1 can be much more efficient than the OLS slope. The efficiency of the Theil estimate to the OLS slope is the same as that for the Hodges-Lehmann estimator in comparison to the mean (Sen, 1968), as the Theil estimate is one of the class of Hodges-Lehmann estimators. The Kendall-Theil line has the desirable properties of a nonparametric estimator: almost as "good" (efficient) as the parametric estimator when all assumptions of normality are met, and much better when those assumptions are not met.

One commonly-asked question is "how much of a departure from a normal distribution is necessary before a nonparametric test has an advantage over its parametric counterpart?". In the case of the Theil and OLS slope estimates, how non-normal must residuals be before the Theil estimate should be used? Are there advantages even in cases where the departure from normality is so small that visual inspection of the data distribution, or formal tests of normality, are unlikely to provide evidence for the lack of normality? Hirsch et al. (1991) tested the two slope estimators under one type of departure from normality, a mixture of two normal distributions. The predominant distribution had a mean of 10 and a standard deviation of 1; the second distribution had a mean of 11 and a standard deviation of 3. Figure 10.2 displays the two individual distributions and figure 10.3 displays a mixture of 95 percent from the first distribution and 5 percent from the second. Visual examination of figure 10.3 reveals only the slightest departure from symmetry. Given sampling variability that would exist in an actual data set it would be exceedingly unlikely that samples from this distribution would be identified as non-normal. Figure 10.4 displays a more substantial departure from normality, a mixture of 80 percent of the first distribution and 20 percent of the second. There is a difference in the shape of the two tails of the distribution, but again the non-normality is not highly noticeable.

Random samples were generated from each of several different mixture distributions containing between 0 and 20 percent of the second distribution. Data from each mixture were treated as a separate response variable in a regression versus a random order x . The true population slope is therefore zero. Both OLS and the Theil slope estimators were computed, and their standard deviations around zero recorded as root mean square error (RMSE). The results are given in figure 10.5 as the ratio of RMSE for the Theil estimator to the RMSE of the regression estimator (Hirsch et al., 1991). A value larger than 1 shows an advantage to OLS; smaller than 1 indicates the Theil estimate to be superior. For the larger sample size ($n=36$) the OLS estimator was more efficient (by less than 10 percent) when the data are not mixed and

therefore normal. With even small amounts of mixtures the Theil estimator quickly becomes more efficient. At a 20 percent mixture the Theil estimator was almost 20 percent more efficient. When the sample size was very small ($n=6$, smaller than typically used in a case study), efficiencies of the two methods were virtually identical.

These results reinforce that when the data or their transforms exhibit a linear pattern, constant variance and near-normality of residuals, the two methods will give nearly identical results. The advantages of familiarity and availability of diagnostics, etc. favor using OLS regression. However, when residuals are not normally distributed, and especially when they contain outliers, the Kendall method will produce a line with greater efficiency (lower variability and bias) than does OLS. Only small departures from normality (not always sufficient to detect with a test or histogram of residuals) favor using a robust approach. Certainly one should check all outliers for error, as discussed in Chapter 1. Do these represent a condition different from the rest of the data? If so, they may be the most important points in the data set. Perhaps another transformation will make the data more linear and residuals near-normal. But outliers cannot automatically be deleted, and often no error can be found. Robust methods like Kendalls or weighted least squares (discussed in sections 10.3 and 10.4) provide protection against disproportionate influence by these distinctive, but perhaps perfectly valid, data points.

For analysis of a small number of data sets, detailed searches for transformations to meet the assumptions of OLS are feasible. OLS is particularly informative in more complex applications requiring incorporation of exogenous effects using multiple regression (see Chapter 11). Cases aren't unusual, however, where no power transformations can produce near-normality due to heavy tails of the distribution. Perhaps the two greatest uses for Kendall's robust fit are 1) in a large study where multiple variables are tested for linear fits at multiple locations without the capability for exhaustive checking of distributional assumptions or evaluations of the sensitivity of results to outliers, and 2) by practitioners not trained in residuals plots and use of transformations to stabilize skewness and heteroscedasticity. A third use is for fitting lines to data which one does not wish to transform.

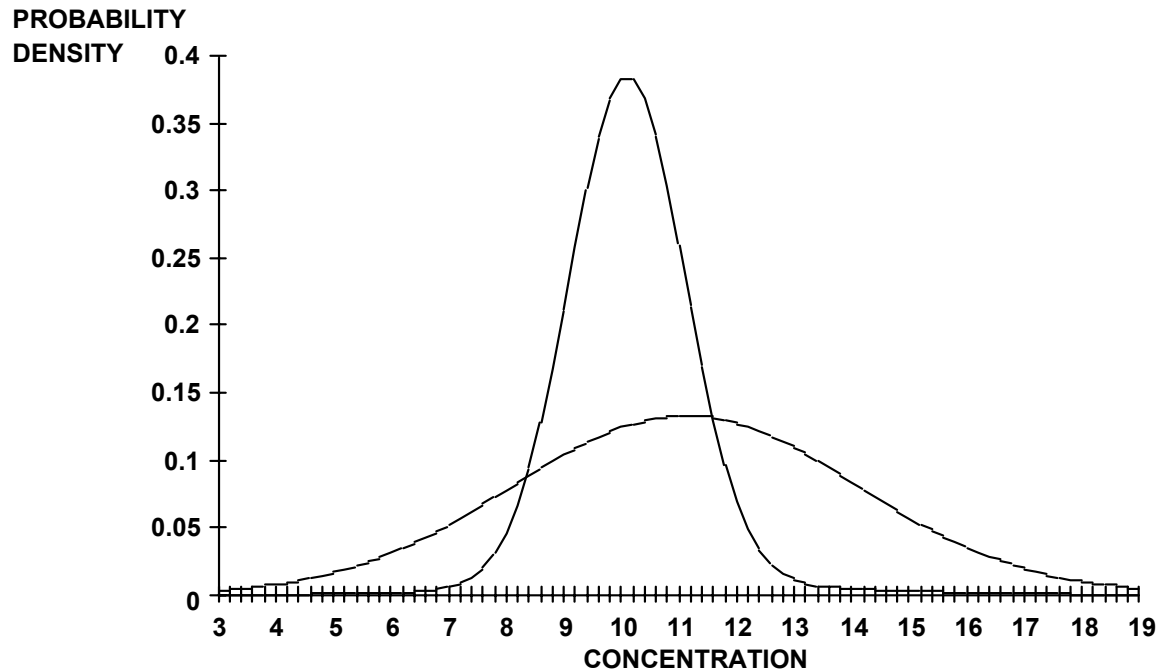


Figure 10.2. Two normal distributions, the first with mean = 10 and standard deviation = 1; the second with mean = 11 and standard deviation = 3 (from Hirsch et al., 1991).

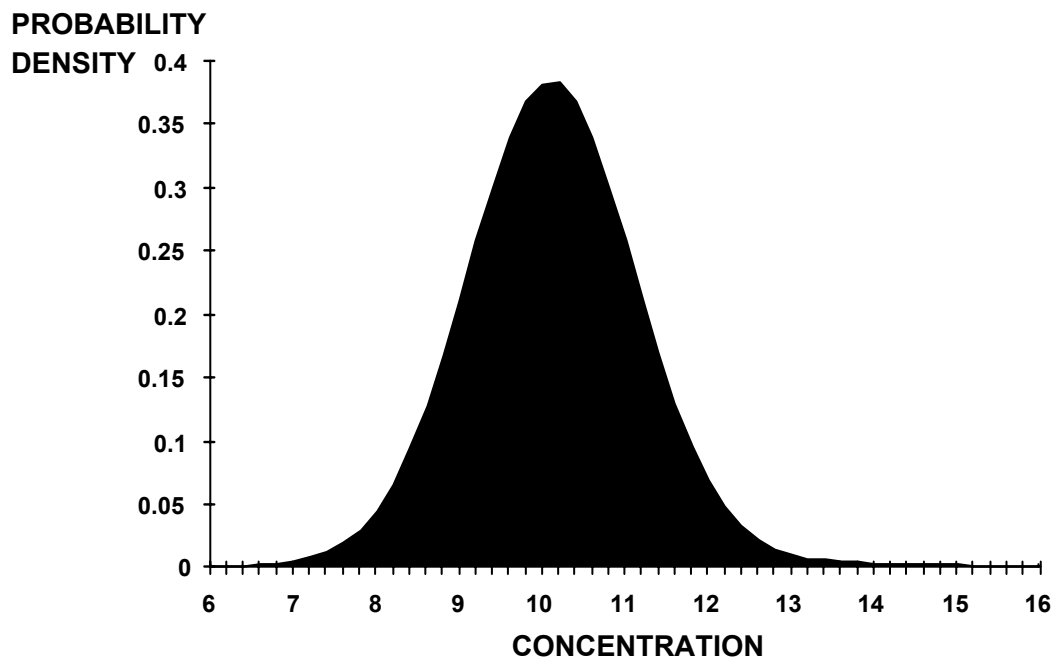


Figure 10.3. A mixture of data from distribution 1 (95 percent) and distribution 2 (5 percent) shown in figure 10.2 (from Hirsch et al., 1991).

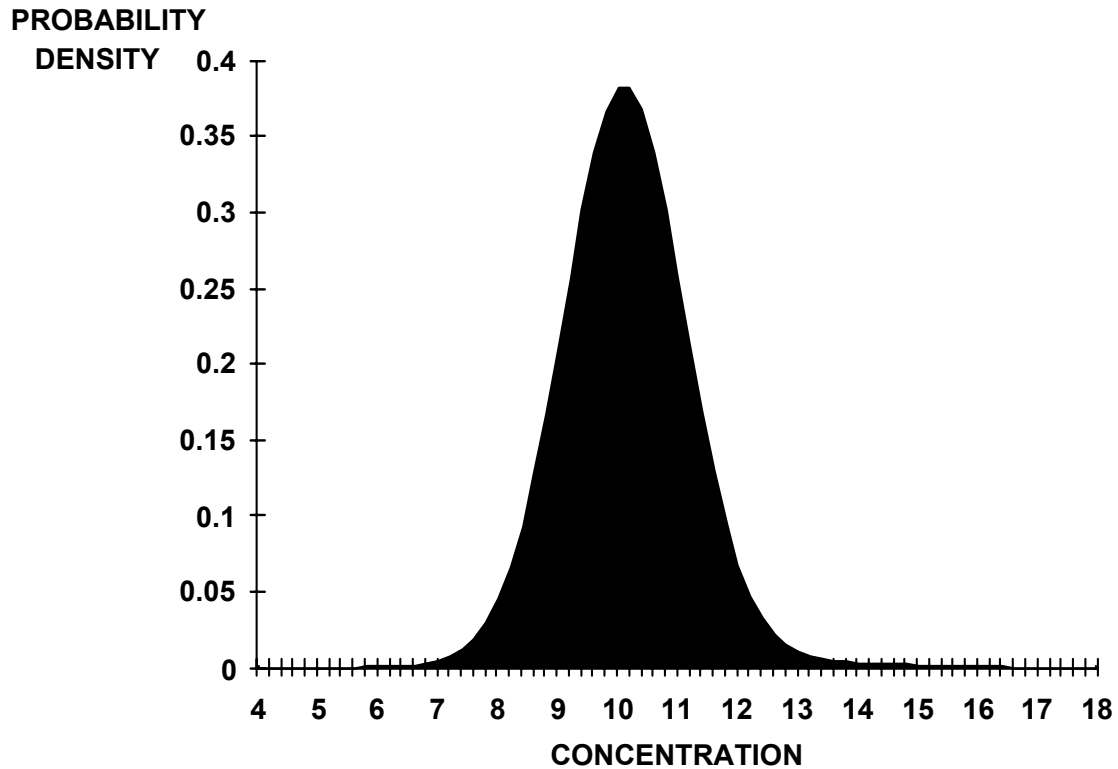


Figure 10.4. A mixture of data from distribution 1 (80 percent) and from distribution 2 (20 percent) shown in figure 10.2 (from Hirsch et al., 1991).

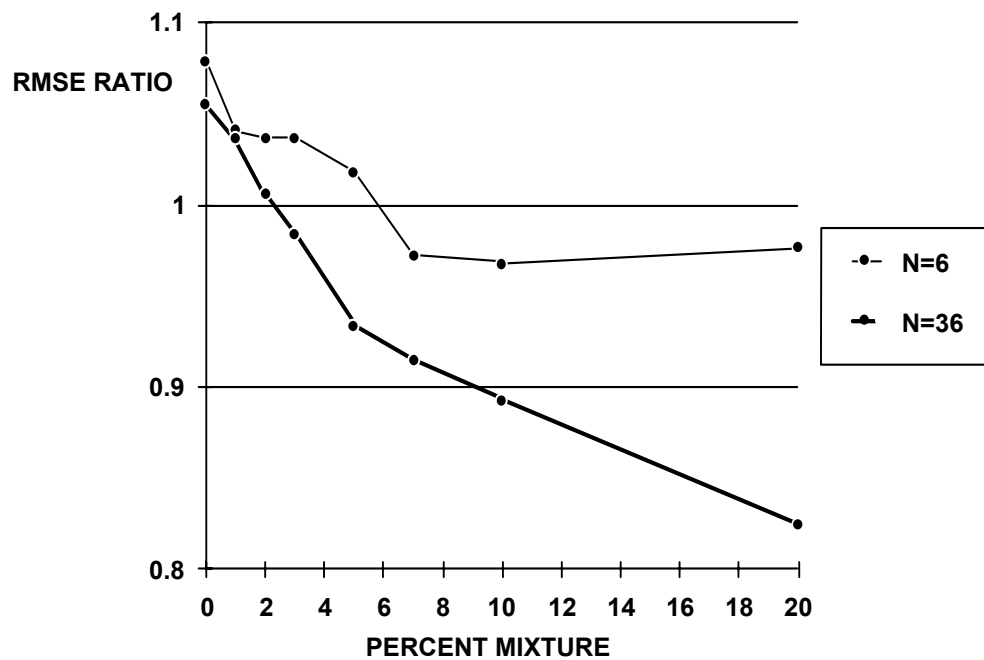


Figure 10.5. Relative efficiency of the Theil slope estimator as compared with the OLS slope.

Efficiency is the ratio of the Theil RMSE to the OLS RMSE, expressed as a function of population mixture and record length (from Hirsch et al., 1991).

Example 2

Figure 10.6 shows an OLS and Kendall-Theil fit to trends in total phosphorus concentrations from 1975 to 1989 in the St. Louis R. at Scanlon, MN. The outliers are accurate values from floods, and therefore cannot be ignored or deleted. The question is whether there is a significant linear trend in concentration over this 14 year period. Here linear fits of concentration versus time are used to test for trend (see Chapter 12 for more on trend tests). The OLS slope is affected by the outliers present. Although the magnitude of the OLS estimate is similar to the Theil slope, the OLS slope does not test as significantly different from zero ($p=0.43$). This is due to inflation of the standard error by outliers in violation of the assumed normality of residuals. The Theil slope is highly significantly different from zero ($p<0.0001$). The Kendall-Theil line is not dependent on assumptions of normality which the data strongly violate.

10.1.3 Test of Significance

The test for significance of the Kendall-Theil linear relationship is the test for $H_0: \tau = 0$. This involves computation of Kendall's S statistic (equation 8.1 of Chapter 8). For $n>10$, the large sample approximation (equation 8.3 of Chapter 8) may be used. The Theil slope estimator \hat{b}_1 is closely related to Kendall's S and τ in the following ways.

1. S is the sum of the algebraic signs of the possible pairwise slopes.
2. If the amount $(\hat{b}_1 X)$ is subtracted from every Y value, the new Y values will have an S and τ very close to zero, indicating no correlation.

If X is a measure of time, as it is for a trend test, subtracting $(\hat{b}_1 X)$ yields a trend-free version of the Y data set.

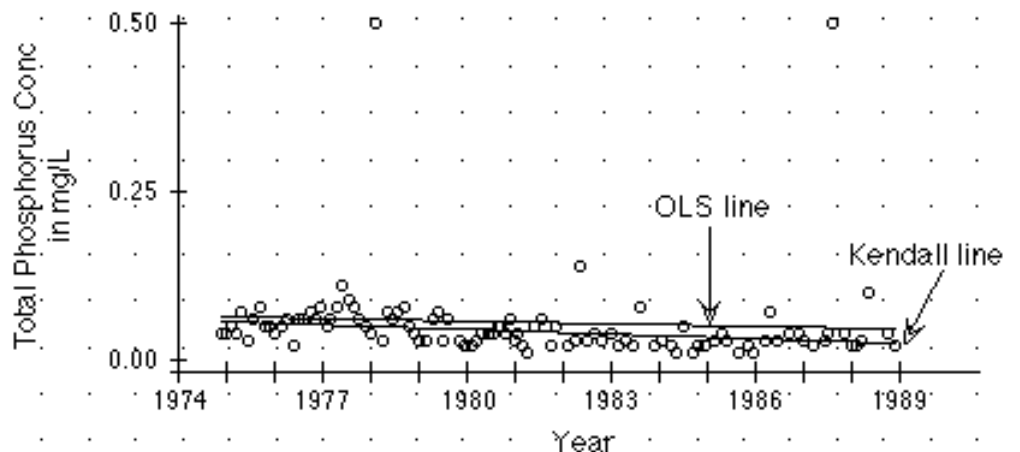


Figure 10.6. Total phosphorus concentrations with OLS and Kendall-Theil fitted lines for the St. Louis River at Scanlon, MN, 1975-1989.

Example 1, cont.

For the example 1 data set, the test of significance is computed as follows. S equals the sum of the signs of pairwise slopes already computed. There are $n(n-1)/2 = 21$ slopes, 20 of which are positive and 1 negative, so that $S = 20 - 1 = 19$. $\text{Tau} = 19/21 = 0.90$. Using table B8 of the Appendix due to the small sample size, the exact two-sided p -value for an S of 19 and $n=7$ is $2 \cdot 0.0014 = 0.003$. (Inappropriately using the large sample approximation for such a small data set, the approximate p -value is 0.007.) Thus Y is significantly related to X in a linear fashion.

10.1.4 Confidence Interval for Theil Slope

Confidence intervals may be computed for the Theil slope \hat{b}_1 with procedures parallel to those used for other Hodges-Lehmann type estimators of earlier chapters. As before, the tabled distribution of the test statistic, in this case table B8 for the exact Kendall's test statistic or a table of standard normal quantiles for the large-sample approximation, is entered to find upper and lower limits corresponding to critical values at one-half the desired alpha level. These critical values are transformed into the ranks corresponding to data points at the ends of the confidence interval.

For small sample sizes, table B8 is entered to find the critical value X_u having a p -value nearest to $\alpha/2$. This critical value is then used to compute the ranks R_u and R_l corresponding to the slope values at the upper and lower confidence limits for \hat{b}_1 . These limits are the R_j th ranked data points going in from either end of the sorted list of $N = n \cdot (n-1)/2$ pairwise slopes. The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data.

$$R_u = \frac{(N + X_u)}{2} \quad [10.3]$$

$$R_l = \frac{(N - X_u)}{2} + 1 \quad [10.4]$$

Example 1, cont.

The $N=21$ possible pairwise slopes between the $n=7$ data pairs for example 1 were:

-9, +1, +1, +1 +1 +1 +1 +1 +1 +1 +1
+1 +1 +1 +1 +1 +3 +3.5 +4.3 +6 +11.

\hat{b}_1 was the median or 11th largest slope. To determine a confidence interval for \hat{b}_1 with $\alpha \cong 0.05$, the tabled critical value X_u nearest to $\alpha/2 = 0.025$ is found to be 15 ($p=0.015$). The rank R_u of the pairwise slope corresponding to the upper confidence limit is therefore

$$R_u = \frac{(21 + 15)}{2} = 18 \quad \text{for } N=21 \text{ and } X_u=15.$$

The rank R_l of the pairwise slope corresponding to the lower confidence limit is

$$R_l = \frac{(21 - 15)}{2} + 1 = 4.$$

So an $\alpha = 2 \cdot 0.015 = 0.03$ confidence limit for \hat{b}_1 is the interval between the 4th and 18th ranked pairwise slope (the 4th slope in from either end), or

$$+1 \leq \hat{b}_1 \leq +3.5.$$

The asymmetry around the estimate $\hat{b}_1 = 1$ reflects the low probability that the slope is less than 1, based on the data.

When the large-sample approximation is used, the critical value $z_{\alpha/2}$ from a table of standard normal quantiles determines the upper and lower ranks of the pairwise slopes corresponding to the ends of the confidence interval. Those ranks are

$$R_u = \frac{N + z_{\alpha/2} \sqrt{\frac{n(n-1)(2n+5)}{18}}}{2} + 1 \quad [10.5]$$

$$R_l = \frac{N - z_{\alpha/2} \sqrt{\frac{n(n-1)(2n+5)}{18}}}{2} \quad [10.6]$$

As an example, for $n=20$ pairs of data there would be $N=(20)(19)/2 = 190$ possible pairwise slopes. \hat{b}_1 is the average of the 95th and 96th ranked slopes. For a 95 percent confidence interval on \hat{b}_1 , $z_{\alpha/2} = 1.96$ and

$$R_u = \frac{190 + 1.96 \cdot \sqrt{950}}{2} + 1 = 126.2$$

$$R_l = \frac{190 - 1.96 \cdot \sqrt{950}}{2} = 64.8$$

the 64.8th ranked slope from either end. Rounding to the nearest integer, the 126th and 65th ranked slopes are used as the ends of the $\alpha=0.05$ confidence limit on \hat{b}_1 .

Further discussion of these equations is in Hollander and Wolfe (1973), pp. 207-208.

10.2 Alternative Parametric Linear Equations

Hirsch and Gilroy (1984) described additional methods for fitting straight lines to data whose slopes and intercepts are computed using moment statistics. These lines differ from the OLS line of Chapter 9, and are more appropriate than that line for certain situations. For example, when X is to be predicted from Y using OLS, the resulting line differs from the OLS line predicting Y from X . This has implications for calibration. When many predictions are to be made and the distribution of those predictions is important (percentiles or spreads are of interest, as well as the mean), the Line of Organic Correlation (LOC) should be used instead of OLS. When describing a functional relationship between two variables without trying to predict

one from the other, LOC is again more appropriate than OLS. When some geographic trajectory is to be computed, the Least Normal Squares (LNS) line should be used.

10.2.1 OLS of X on Y

The OLS regression of Chapter 9 considered the situation where a response variable Y was to be modeled, enabling estimates of Y to be predicted from values of an explanatory variable X. Estimates of slope and intercept for the equation were obtained by minimizing the sum of squares of residuals in units of Y. Thus its purpose was to minimize errors in the Y direction only, without regard to errors in the X direction. The equation may be written as:

$$Y_i = \bar{Y} + r \frac{s_y}{s_x} (X_i - \bar{X}) \quad [10.7]$$

where r is Pearson's linear correlation coefficient, s_y and s_x are the standard deviations of the Y and X variables, and $(r s_y/s_x) = (r \sqrt{SS_y}/\sqrt{SS_x}) = b_1$, the OLS estimate of slope (see Chapter 9). Assuming the linear form of the model is correct and that X and Y are measured without error, OLS will lead to estimates of Y_i for any given X_i which are unbiased and have minimum variance. This means that OLS is the preferred method of estimating a single value of Y given a value of X, where X is measured without error.

In contrast, situations occur where it is just as likely that X should be predicted from Y, or that the two variables are equivalent in function. One classic example is in geomorphology, where relations between the depth and width of a stream channel are to be related. It is as reasonable to perform a regression of depth on width as it is of width on depth. A second example is the relation between dissolved solids concentration and "residue on evaporation" or ROE, an alternate measure of the amount of dissolved material in a water sample. Either could be chosen to model as a function of the other, and usually a description of their relationship is what is of most interest.

It is easy to show, however, that the two possible OLS lines (Y on X and X on Y) differ in slope and intercept. Following equation [10.7], reversing the usual order and setting X as the response variable, the resulting OLS equation will be

$$X_i = \bar{X} + r \frac{s_x}{s_y} (Y_i - \bar{Y}) \quad [10.8]$$

which when solved for Y becomes

$$Y_i = \bar{Y} + \frac{1}{r} \frac{s_y}{s_x} (X_i - \bar{X}) \quad [10.9]$$

Let $b_1' = (1/r \cdot s_y/s_x)$, the slope of X on Y re-expressed to compare with slope b_1 . Contrasting [10.7] and [10.9], the slope coefficients $b_1 \neq b_1'$. Thus the two regression lines will differ unless

the correlation coefficient r equals 1.0. In figure 10.7, these two regression lines are plotted for the dissolved solids and ROE data of Appendix C12.

The choice of which, if either, of the OLS lines to use follows a basic guideline. If one is to be predicted from the other, the predicted variable should be assigned as the response variable Y . Errors in this variable are being minimized by OLS. However, when only a single line describing the functional relationship between the two variables is of interest, neither OLS line is the appropriate approach. Neither OLS line uniquely or adequately describes that relationship. A different linear model having a unique solution should be used instead -- the line of organic correlation.

10.2.2 Line of Organic Correlation

The line of organic correlation (LOC) was proposed as a linear fitting procedure in hydrology by Kritskiy and Menkel (1968) and applied to geomorphology by Doornkamp and King (1971). Its theoretical properties were discussed by Kruskal (1953). The line also has been called the "geometric mean functional regression" (Halfon, 1985), the "reduced major axis" (Kermack and Haldane, 1950), the "allometric relation" (Teisser, 1948) and "Maintenance of Variance - Extension" or MOVE (Hirsch, 1982). It possesses three characteristics preferable to OLS in specific situations:

- LOC minimizes errors in both X and Y directions.
- It provides a unique line identical regardless of which variable, X or Y , is used as the response variable, and
- The cumulative distribution function of the predictions, including the variance and probabilities of extreme events such as floods and droughts, estimates those of the actual records they are generated to represent.

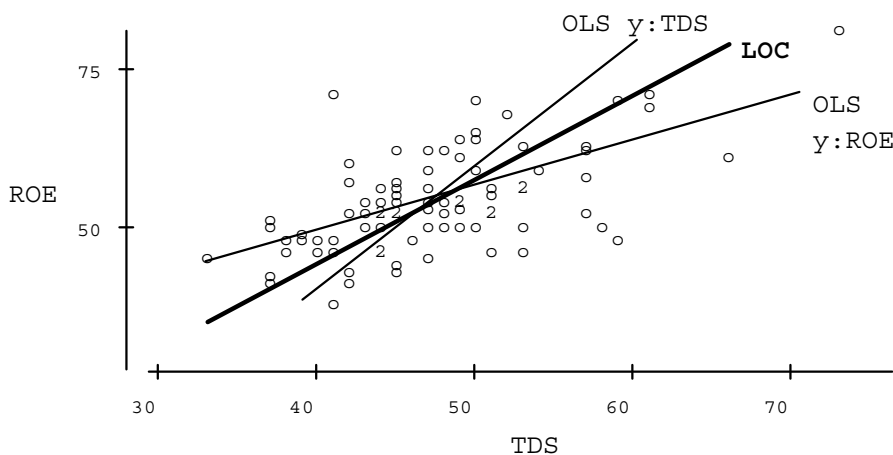


Figure 10.7 Three straight lines fit to the same data.

The LOC minimizes the sum of the areas of right triangles formed by horizontal and vertical lines extending from observations to the fitted line (figure 10.8). By minimizing errors in both directions it lies between the two OLS lines on a plot of Y versus X (see figure 10.7). The slope of the LOC line equals the geometric mean of the Y on X and X on Y OLS slopes:

$$b_1'' = \sqrt{b_1 b_1'} = \text{sign}[r] \cdot \frac{s_Y}{s_X}$$

where b_1'' is the slope of the LOC line

$$Y_i = b_0'' + \text{sign}[r] \cdot \frac{s_Y}{s_X} \cdot X_i \quad [10.10]$$

So the correlation coefficient in the equation for OLS slope is replaced by the algebraic sign (+ or -) of the correlation coefficient with LOC. The magnitude of the LOC slope b_1'' is determined solely by the ratio of standard deviations s_Y/s_X . Performing LOC of X on Y will give the identical line as does the LOC of Y on X.

LOC is therefore used for two purposes, corresponding to the three above attributes:

- a , b) to model the correct functional relationship between two variables, both of which are measured with error.
- c) to produce a series of estimates \hat{Y}_i from observed X_i whose distributional properties are similar to those expected had the Y_i been measured. Such estimates are important when the probability distribution (variance or percentiles) of the estimates, and not just the mean or an individual estimate, are to be interpreted and used.

Examples of the first use for LOC include the geomorphic relationships cited above, describing the relation between bioaccumulation and octanol-water partition coefficients (Halfon, 1985), or other applications where the slope is to take on physical meaning rather than interest in prediction of values of one variable.

One example of the second use for LOC is the extension or fill-in of missing observations. This use for record extension has been the major application of LOC to water resources thus far. As an example, suppose two nearby sites overlap in their gaged record. The streamflow for the site with the shorter record is related to that at the longer (the "base") site during the overlap period. Using this relationship, a series of streamflow data at the shorter site is estimated during an ungaged period based on flows at the base site. If the OLS equation were used to estimate streamflows, the variance of the resulting estimates would be smaller by a factor of R^2 than it should be. OLS reduces the variance of estimates because the OLS slope is a function not only of the ratio of the standard deviations s_Y/s_X , but also of the magnitude of the correlation coefficient r . Only when $|r| = 1$ do OLS estimates possess the same variance as would be expected based on the ratio of variances for the original data. To see this more clearly, take the extreme case where $r=0$, and there is no relationship between

Y and X. The slope then equals 0, and all OLS estimates would be identical and equal to \bar{Y} . The variance of the estimates is also zero. As R^2 decreases from 1 to 0, the variance of OLS estimates is proportionately reduced. This variance reduction is eliminated from LOC by eliminating the correlation coefficient from the equation for slope. The estimates resulting from the LOC have a variance in proportion to the ratio of the variances s_y^2/s_x^2 from the original data.

When multiple estimates are to be generated and statements made about probabilities of exceedance, such as flood-flow probabilities, probabilities of low-flows below a water supply intake, or probabilities of exceeding some water-quality standard, inferences are made which depend on the probability distribution of the estimated data. In these cases LOC, rather than OLS, should be used to generate data. OLS estimates would substantially underestimate the variance because they do not include the variability of individual values around the regression line (Hirsch, 1982). As a consequence, the frequency of extreme events such as floods, droughts, or exceedance of standards would be underestimated by OLS.

Variations on using LOC for hydrologic record extension have been published by Vogel and Stedinger (1985) and Grygier et al. (1989).

All three of the lines discussed thus far have two identical characteristics. They are invariant to scale changes, so that changing the Y or X scale (from English to metric units, for example) will not change the estimates of slope or intercept after re-expressing them back into their original scales. However, if the X and Y axes are rotated and lines re-computed, the second set of estimates will differ from the first following re-expression into the original orientation. This second property is not desirable when the original axes are of arbitrary orientation, such as for latitude and longitude. The line discussed in the next section can be fit when invariance to spatial orientation is desired.

10.2.3 Least Normal Squares

Least normal squares is the line which minimizes the squared distances between observed points and the line, where distances are measured perpendicular (normal) to the line. The slope can be expressed as in figure 10.8

$$b = -A + \frac{\sqrt{r^2 + A^2}}{r} \quad \text{where } A = \frac{1}{2} \left(\frac{s_x}{s_y} - \frac{s_y}{s_x} \right) \quad [10.11]$$

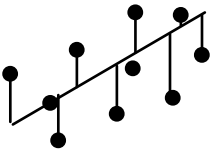
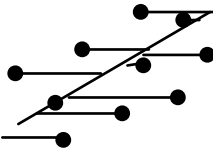
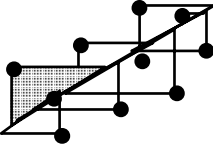
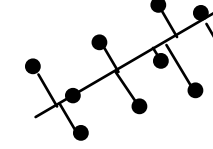
<u>Method</u>	<u>Minimizes:</u>	<u>Slope</u>	<u>Scale Change</u>	<u>Rotation</u>
OLS Y on X		$b_1 = r \frac{s_y}{s_x}$	invariant	changes
OLS X on Y		$b_1' = \frac{1}{r} \frac{s_y}{s_x}$	invariant	changes
LOC		$b_1'' = \text{sign}[r] \frac{s_y}{s_x}$	invariant	changes
LNS		$b = \frac{-A + \sqrt{r^2 + A^2}}{r}$ where $A = 0.5 \left(\frac{s_x}{s_y} - \frac{s_y}{s_x} \right)$	changes	invariant

Figure 10.8 Characteristics of four parametric methods to fit straight lines to data

An appealing property of LNS is its invariance to rotation of axes. This is desirable when the coordinate system in which the data are measured is arbitrary. The most common example of this is where X and Y are physical locations, such as latitude and longitude. If the axes are rotated, the X and Y coordinates of the data recomputed, and the LNS line recomputed, it will coincide exactly with the LNS line for the data prior to rotation. This is not so with OLS or LOC. However, the LNS line is not invariant to scale changes. The LNS line expressed in any scale will differ depending on the scale in which the calculations were made. For example, the LNS line relating concentration in mg/L to streamflow in cubic feet per second will differ from the LNS line for the same data using streamflow in cubic meters per second. This attribute makes LNS poorly suited to describe the relation between most common water resources variables. Where LNS is appropriate is in computing trajectories minimizing distances between observed points in space. Kirby (1974) thus used LNS to compute the straight line traverse of a ship from a set of coordinate locations taken along its trip.

10.2.4 Summary of the Applicability of OLS, LOC and LNS

To summarize the application of each of the above parametric procedures:

1. To estimate individual values of one variable from another variable, use OLS (assuming the data are linear and homoscedastic). This holds regardless of causality, and regardless of whether there are errors in measurement of the explanatory variable.
2. To estimate multiple values of one variable from another variable in order to make statements about the probability distribution, use LOC. This preserves the characteristics of the entire distribution, avoiding the downward bias in variance of the OLS estimates.
3. To describe the functional relationship between two variables with the primary interest in the slope coefficient, use LOC.
4. To determine the geographic trajectory which minimizes the differences from observed data, use LNS.

10.3 Weighted Least Squares

Data may exhibit a linear pattern yet have non-constant variance (heteroscedasticity -- see figure 10.9). Corrections for non-constant variance involving a power transformation will often alter the linear pattern to one which is curved. Also, transformation into differing units may not be desirable, due to retransformation bias of the estimates (see Chapter 9). Finally, the data may have known inherent differences in their variances, such as when means or other summary statistics based on unequal-sized data sets are used as the explanatory variable. When the constant variance assumption of OLS is violated, an alternate method called weighted least squares (WLS) should instead be employed.

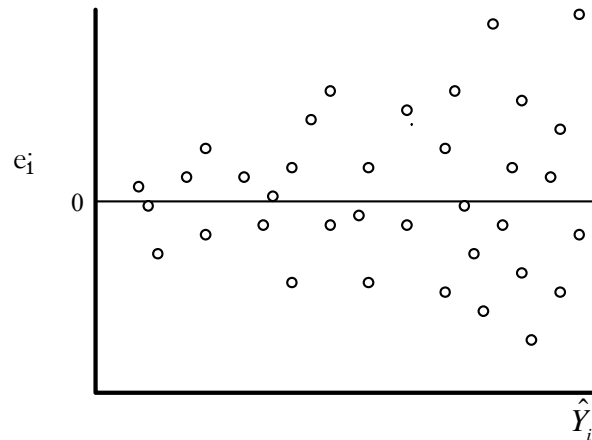


Figure 10.9 Heteroscedastic data.

With WLS, each squared residual $(Y_i - \hat{Y}_i)^2$ is weighted by some weight factor w_i in such a way that observations with greater variance have lesser weight. Thus "less reliable" observations have less influence on the resulting linear equation than "more reliable" observations. The fitted WLS equation minimizes the squares of the weighted residuals. To evaluate whether this weighting has corrected for heteroscedasticity, a weighted residuals plot should be drawn. In this plot the weighted residuals, $e_i \sqrt{w_i}$ are plotted versus $\hat{Y}_i \sqrt{w_i}$. The pattern of weighted residuals can be interpreted as with any other residuals plot.

One common use of WLS in water resources arises when basin characteristics are used to estimate flood percentiles (Tasker, 1980). For example, estimates of the 100-year flood at ungaged sites can be made from a log-log regression of sample estimates of 100-year floods for gaged sites within a region versus drainage area. The flood flows used to construct the regression will have differing variances for different sites, depending on their record lengths n . Sample estimates based on longer records are more reliable, and will have lower variance, than for stations with less data. Therefore estimates from longer records should be given a stronger effect on the regression line. If all original observations are assumed to have constant variance σ^2 , then the weights w_i for the weighted regression will be proportional to the record lengths n_i at each station.

Further weighting could reflect any spatial correlation between the sites. This is called generalized least squares, and is applied to hydrology by Stedinger and Tasker (1985). An example of weighting in response to differential sampling within a stratified sampling design is given by DuMouchel and Duncan (1983).

A more empirical method of weighting occurs by setting weights inversely proportional to the sample variance of the response variable at that location. This variance is rarely known ahead of time, so that weights are computed based on residuals from an ordinary least squares regression (OLS) in the following manner:

- 1) OLS regression is computed for Y versus X . Residuals are plotted against \hat{Y} , and nonconstant variance is seen.
- 2) Observations with similar X 's are grouped, and the variance of the observations in each group s_y^2 is calculated. These variances are plotted versus X_i for each group.
- 3) Assign s_y^2 to each observation in group i . Weights $w_i = 1/s_y^2$.

Weighted least squares can be computed using software for unweighted multiple regression by employing a data transformation $Y_i' = c_i Y_i$, where each observation Y_i is multiplied by the square root of the weight for that point ($c_i = \sqrt{w_i} = 1/s_y$). The X_i must also be weighted as $X_i' = c_i X_i$. A weighted intercept term must also be included as a new "variable" I_i' , consisting of a vector of c_i 's, one per observation. The transformed Y_i' are then related by multiple regression to X_i' and I_i' using the "no intercept" option (the I' column is the weighted intercept). The resulting coefficients are the coefficients of the weighted least squares line.

Example 3

Total dissolved solids (TDS) from Appendix C12 are plotted versus time, and an increasing variance is seen (figure 10.10). Regression of TDS versus time produces:

$$\text{TDS} = -1627 + 0.844 \cdot \text{Time}, \quad t\text{-statistic} = 4.62 \quad p = <0.001$$

where Time is in years. A residuals plot would also show increasing variance.

However, this equation puts undue emphasis on the more recent data, which have the largest variability. The variability seems to increase after 1985, therefore the data are split into two periods, and the variance of TDS is computed separately for each period.

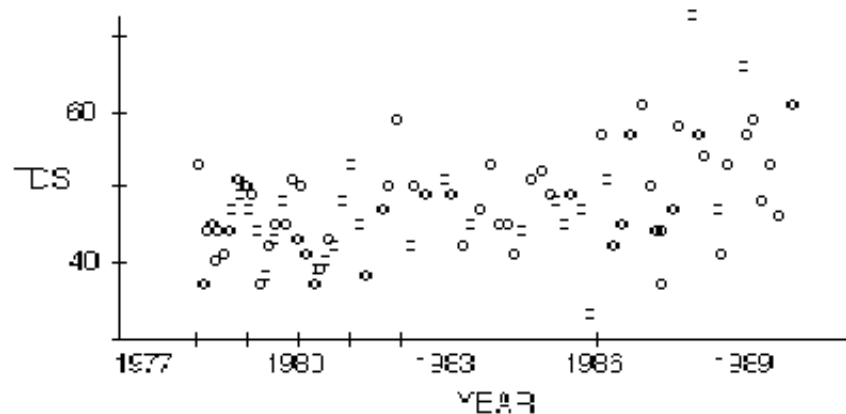


Figure 10.10 TDS data with non-constant variance (heteroscedasticity).

The variance for the pre-1985 period is 24.18, while after 1985 it is 71.80. The reciprocal of these values is assigned as the weight function for each observation in the respective groups, and a weighted least squares regression is performed. This results in:

$$\text{TDS} = -1496 + 0.778 \cdot \text{Time. } t\text{-statistic} = 4.10 \quad p = <0.001$$

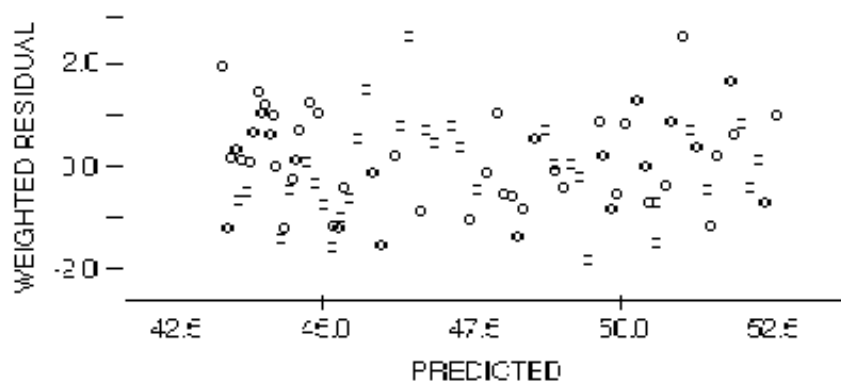


Figure 10.11 Weighted residuals plot of the TDS data.

A plot of the weighted residuals versus predicted values is shown in figure 10.11. The weighted residuals have constant variance. Thus the weighted least squares line should be preferred to the unweighted line, because it more closely conforms to one of the assumptions of least squares regression -- constant variance of residuals.

10.4 Iteratively Weighted Least Squares

OLS regression can be thought of as a "linear mean", with both desirable and undesirable properties similar to a mean. One undesirable property is that outliers can "pull" the location of the line (estimates of slope and intercept) in their direction, much in the same fashion as the sample mean is affected by an outlier. The resulting residuals corresponding to the outlying point may be small, making that point difficult to discern as unusual. Such outliers must be detected using influence statistics (see Chapter 9). In addition to detecting outliers, it may be desirable to limit their influence on the regression line, similar in objective to the Kendall-Theil method given in section 10.1. A second method for doing so, somewhat analogous to a trimmed mean, is a robust regression method called iteratively weighted least squares (IWLS). Unlike Kendall-Theil, IWLS is applicable in the multiple regression context.

The goal of any robust regression is to fit a line not strongly influenced by outliers. This leaves large residuals for the outliers, but a better fit to most other points. IWLS produces models similar to OLS when the underlying residuals distribution is normal, where OLS would have been reasonable to use. Alternate methods of robust regression to IWLS include "least median of squares" and "least absolute value" (Rousseeuw and Leroy, 1987), both of which minimize a more robust measure of error than least squares.

With IWLS, weights are derived from the data. An OLS is first computed -- all weights are initially set equal to one. Points nearest the OLS line are then given weights near one, while points further away have lesser weight. A weighted least squares is computed, and the process repeated. After about two iterations the weights become stabilized, and the final iteratively weighted least squares line results.

There are several weight functions which have been used to compute weights. A common and useful one is the bisquare weight function (Mosteller and Tukey, 1977):

$$w_i = \begin{cases} (1 - u_i^2)^2 & \text{for } |u_i| \leq 1 \\ 0 & \text{for } |u_i| > 1 \end{cases}$$

$$\text{where } u_i = \frac{Y_i - \hat{Y}_i}{c \cdot S}$$

c = constant, and

S = some robust measure

of spread of the residuals $(Y_i - \hat{Y}_i)$.

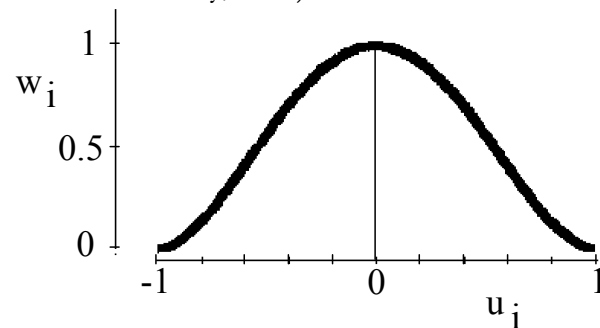


Figure 10.12 Bisquare Weight Function

The purpose of the divisor $c \cdot S$ is to make u_i invariant to scale changes.

Common choices for c and S are

- $c = 3$ and $S =$ the IQR of the residuals. For a normal distribution $\text{IQR} \cong 4/3 \sigma$, so that when $c = 3$, $c \cdot S \cong 4\sigma$. This is a margin sufficiently wide to include most or all observations when the distribution is near-normal, and yet protect against outliers when the distribution is markedly non-normal.
- $c = 6$ and $S =$ the MAD, the median absolute deviation from the previous line, or median $|\text{residuals}|$. Again $c \cdot S \cong 4\sigma$ ($\text{MAD} \cong 2/3 \sigma$ for a normal distribution).

Note that since the sample standard deviation is strongly distorted by outliers, it would be a poor choice as the measure of spread S . This highlights the failing of all parametric tests for outliers: if the criteria for declaring a value as an outlier is strongly influenced by those same outliers, it will be inflated to the point of declaring too few data as outliers. Either the MAD or IQR are more appropriate than the standard deviation for this purpose.

After calculating the IQR or MAD of residuals from an OLS, the first set of weights are produced. These weights are used in the first weighted least squares, from which new residuals are used to compute new weights. The process is repeated until the weights stabilize -- in most cases only two iterations are required.

Example 3

TCE concentrations were measured in wells from the Upper Glacial Aquifer, Long Island, NY., and related to population density (Eckhardt et al., 1989). Below are listed the percent of wells with TCE concentrations above the detection limit (%DET), by population density of the surrounding land (POPDEN). Compute the robust regression equation (2 iterations) to predict detection percentage from population density.

%DET	0.64	4.80	10.20	22.50	25.00	25.00	67.00	38.00	31.30
POPDEN	1	2	3	5	6	8	9	11	13

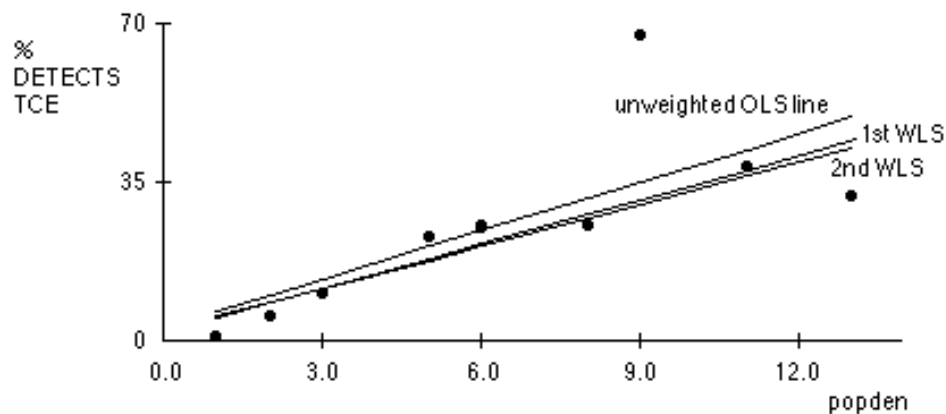


Figure 10.13 TCE concentrations on Long Island (Eckhardt, 1989)

The OLS (unweighted) regression equation is $\%DET = 2.00 + 3.56 \cdot POPDEN$ with a t-statistic of 2.86. This line is pulled up by the one outlier at a population density of 9 which doesn't fit the rest of the data very well (figure 10.13). The residuals e_i from this OLS line are used to establish bisquare weights for the first WLS line.

e_i :	-0.414	-0.345	-0.191	0.199	0.120	-0.404	2.478	-0.253	-1.545
w_i :	0.929	0.945	0.982	0.978	0.992	0.913	0.000	0.971	0.326

The outlying point is sufficiently far from the line that it receives a weight of zero. The first weighted regression equation is then $\%DET = 0.93 + 3.23 \cdot POPDEN$, with a t-statistic of 6.93. This is shown as "1st WLS" in figure 10.13. Again, residuals are computed from this equation, and a new set of weights computed:

w_i :	0.945	0.970	0.999	0.872	0.903	0.986	0.000	0.989	0.489
---------	-------	-------	-------	-------	-------	-------	-------	-------	-------

The 2nd iteration weighted regression equation is then $\%DET = 1.24 + 3.10 \cdot POPDEN$, similar to the previous iteration, with a t-statistic of 6.63. Figure 10.13 shows this line as "2nd WLS". The residual for the outlying point remains large, while the line fits the majority of the data quite well. This is the objective of a robust regression.

10.5 Smoothing

Smoothing differs in purpose and form from the previous methods. It is an exploratory technique, having no simple equation or significance tests associated with it. The most common smooths estimate the center of the data -- the conditional mean or median of Y as X changes. The lack of an equation is a strength in the sense that a smooth is not constrained by some prior assumption as to the mathematical function of the relationship. Rarely are there theoretical grounds for choosing one function over another in modeling Y versus X. For large data sets it is common to visually identify departures from a simple function which could only be modelled by incorporating several high order terms. This can cause instability near or beyond the range of the data. The shape of a smooth is not specified *a priori*, but is determined solely by the data.

Middle smooths allow the data to dictate the location of a smooth curve which goes through the middle of the data. They are used to highlight trends or patterns in the data on a scatterplot. These patterns are often difficult to see. The human eye only poorly follows the central tendency of a scatterplot; the range of data dominates visual impression. Adding a line through the middle draws attention to the center of the plot, aiding judgement of whether the pattern is linear, indicating where breaks in slope occur, etc.

10.5.1 Moving Median Smooths

The simplest smooths are moving averages or medians. Data are smoothed by calculating the mean or median of a portion of the total data within some 'window' of influence around a given

X_0 . This is repeated while setting X_0 equal to nearly every X value in the data set. As before, outliers will influence moving averages (means) more strongly than medians, so that moving averages are more erratic than medians in the vicinity of outliers. Moving medians therefore are more resistant to outliers than are moving averages.

Suppose a 5-point moving median is to be computed. A 'window' of width equal to 5 data points is begun at the left of the X-Y plot. The median of the 5 Y values within the window is computed, and plotted at the center of the window ($X_0 = 3$ rd point from the left) to form the first value of the smooth. Data outside the window have no influence on the smoothed value. The X window is shifted to the right by one data point, a new median of the 2nd through 6th points calculated, and this value plotted at the new $X_0 = 4$ th point from the left. This shifting and computation progressively continues through the final window, composed of the rightmost 5 points. All medians are then connected by straight lines to form the moving-median smooth.

Figure 10.14 shows an 11-point moving median smooth for sand concentrations in the Colorado River at Lees Ferry, Arizona. Moving medians are convenient for hand computation, but produce a "rough" pattern unless the window size is quite large. Large windows result in the undesirable characteristic that data far from X_0 influence the resulting value as much as data nearby. To avoid this, more complex smoothing routines are now performed by computer.

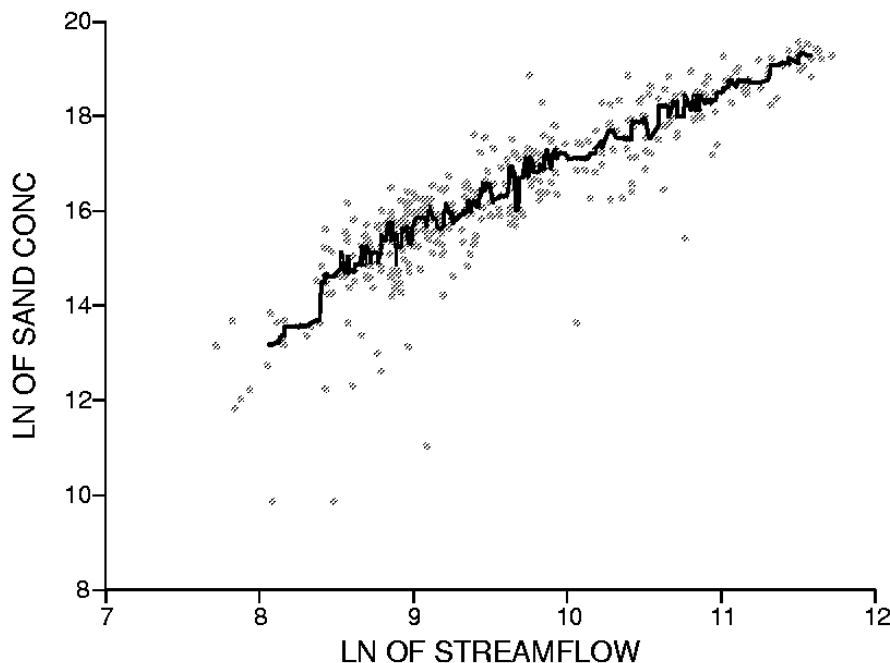


Figure 10.14 11 point moving median of sand concentrations in the Colorado River at Lees Ferry, Arizona, 1949-1970.

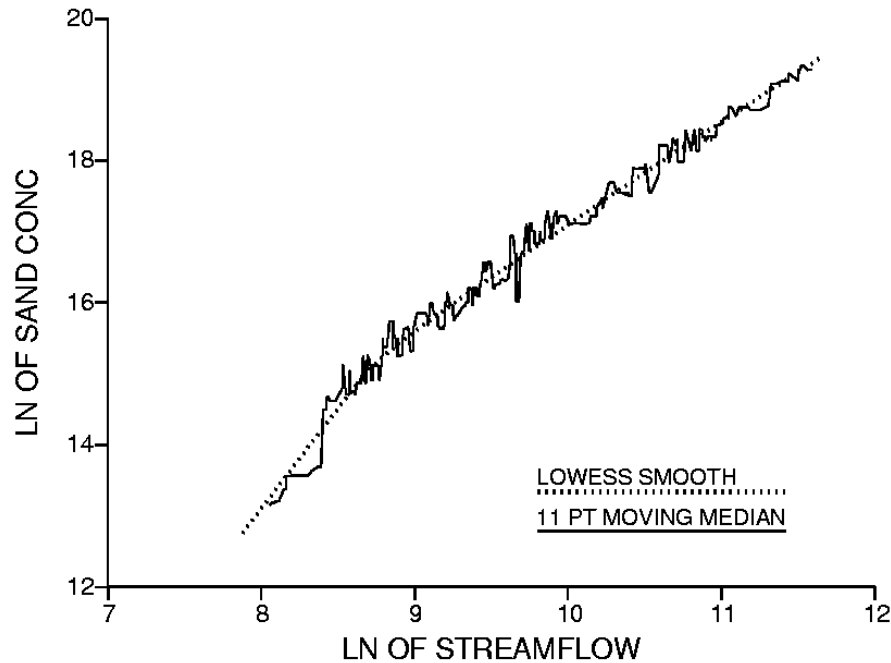


Figure 10.15 11 point moving median and LOWESS smooths of the Lees Ferry data

These allow the data nearer the center of the window to influence the smoothed value more than those further away. They also allow the smoothness of the final fit to be adjusted to the needs of the data analyst. One of the most flexible and useful smoothing algorithms is called LOWESS. In figure 10.15 the 11 point moving median smooth is compared to a LOWESS smooth for the Lees Ferry data.

10.5.2 LOWESS

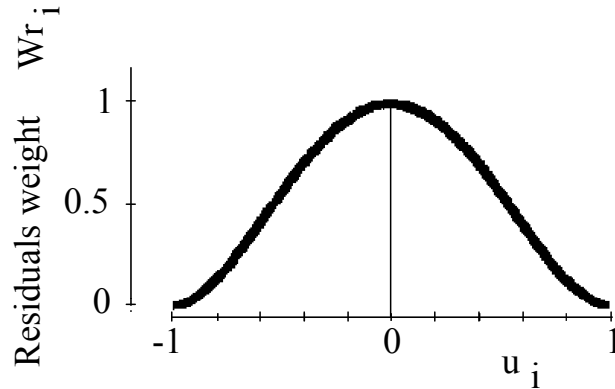
LOWESS, or LOcally WEighted Scatterplot Smoothing (Cleveland et al., 1979) is computationally intensive. It involves fitting at least 2^n weighted least squares equations. At every X_0 , a \hat{Y} is computed from a WLS regression whose weights are a function of both the distance from X_0 and the magnitude of the residual from the previous regression (an iterative procedure). The robust regression weights w_i are computed by

$$w_i = wx_i \cdot wr_i$$

where wx_i , the distance weight, is a function of the distance between the center of the window X_i and all other X . The residuals weight wr_i is a function of $|Y_i - \hat{Y}_i|$, the distance in the Y direction between the observed Y_i and the value predicted from the previous WLS equation. A point will receive a small weight, and therefore have little influence on the smoothed \hat{Y} , if it is either far from the center of the window in the X direction or has a large residual in the Y direction. The measure of how quickly weights decrease as distances increase in the X and Y directions is determined by the weight function. For a point at (X_i, Y_i) , the bisquare weight is determined as

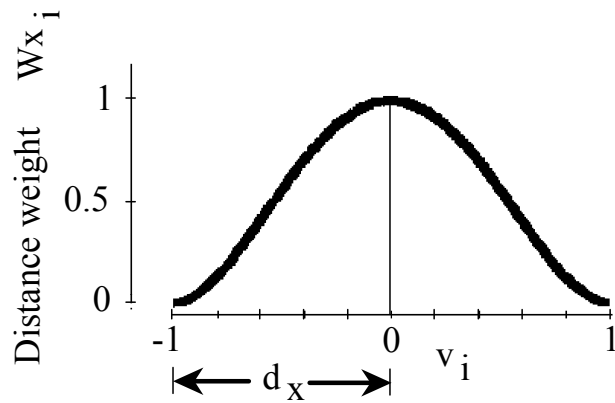
$$wr_i = \begin{cases} (1 - u_i^2)^2 & \text{for } |u_i| \leq 1 \\ 0 & \text{for } |u_i| > 1 \end{cases}$$

$$\text{where } u_i = \frac{Y_i - \hat{Y}_i}{6 \cdot \text{median of all } |Y_i - \hat{Y}_i|}$$



$$wx_i = \begin{cases} (1 - v_i^2)^2 & \text{for } |v_i| \leq 1 \\ 0 & \text{for } |v_i| > 1 \end{cases}$$

$$\text{where } v_i = \frac{X_i - X}{d_x}$$



where d_x = half width of window = m th largest $|X_i - X|$

$$m = Nf$$

N = sample size

f = smoothness factor specified at outset.

Smoothness of LOWESS is varied by altering the window width, as controlled by the smoothness factor f (figure 10.16). As f is increased, the window size is increased, and more points influence the magnitude of \hat{Y} . Selection of an appropriate f is determined subjectively according to the purpose for which the smooth is used.

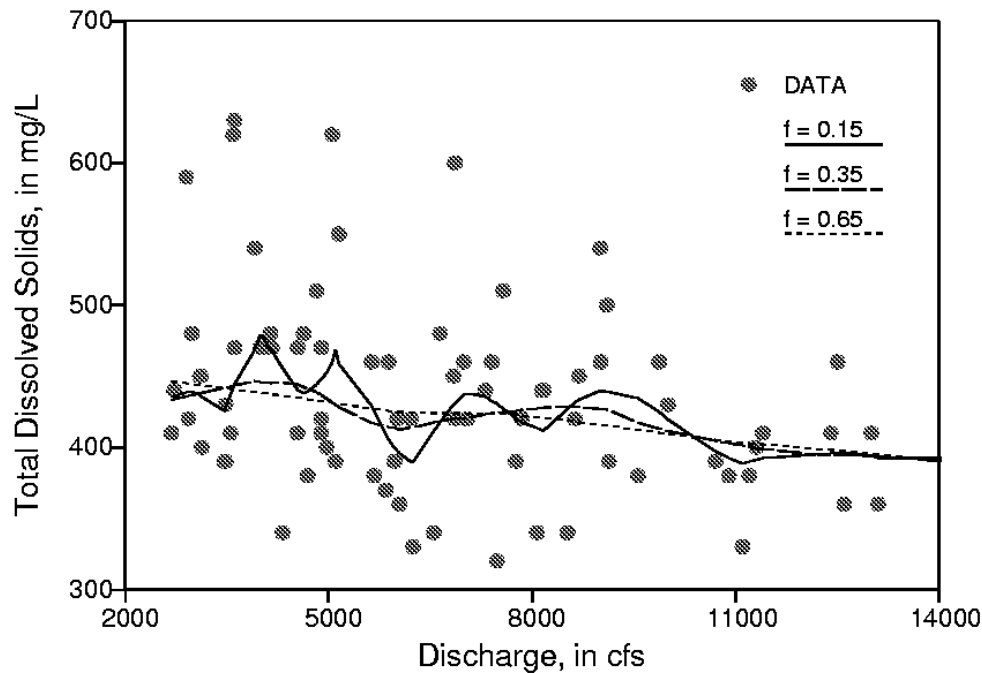


Figure 10.16 Three smooths of the same data with differing smoothness factors f .

Three examples of situations in which LOWESS smooths greatly aid data analysis are:

1. To emphasize the shape of the relationship between two variables on a scatterplot of moderate to large sample size. Adding a line through the middle draws attention to the center of the plot, aiding judgement of how the two variables are related.
2. To compare and contrast multiple large data sets. Plotting all data points with differing symbols per group does not provide the clarity necessary to distinguish similarities and differences between groups. Instead, computing and plotting LOWESS smooths without the data may give great insight into group characteristics. For example, Welch et al. (1988) used LOWESS to describe the relationship between arsenic and pH in four physiographic regions of the Western United States (figure 2.26 in Chapter 2). Thousands of data points were involved; a scatterplot would have shown nothing but a blob of data. The smooths clearly illustrated that in three regions arsenic concentrations increased with increasing pH, while in the fourth no increase was observed. Smooths were also used by Schertz and Hirsch (1985) to illustrate regional patterns in atmospheric precipitation chemistry. They used one smooth per station to display simultaneous changes in sulfate and other chemical concentrations occurring over broad regions of the country (figure 10.17). These relationships would have gone unnoticed using scatterplots -- the underlying patterns would have been obscured by the proliferation and scatter of the data.

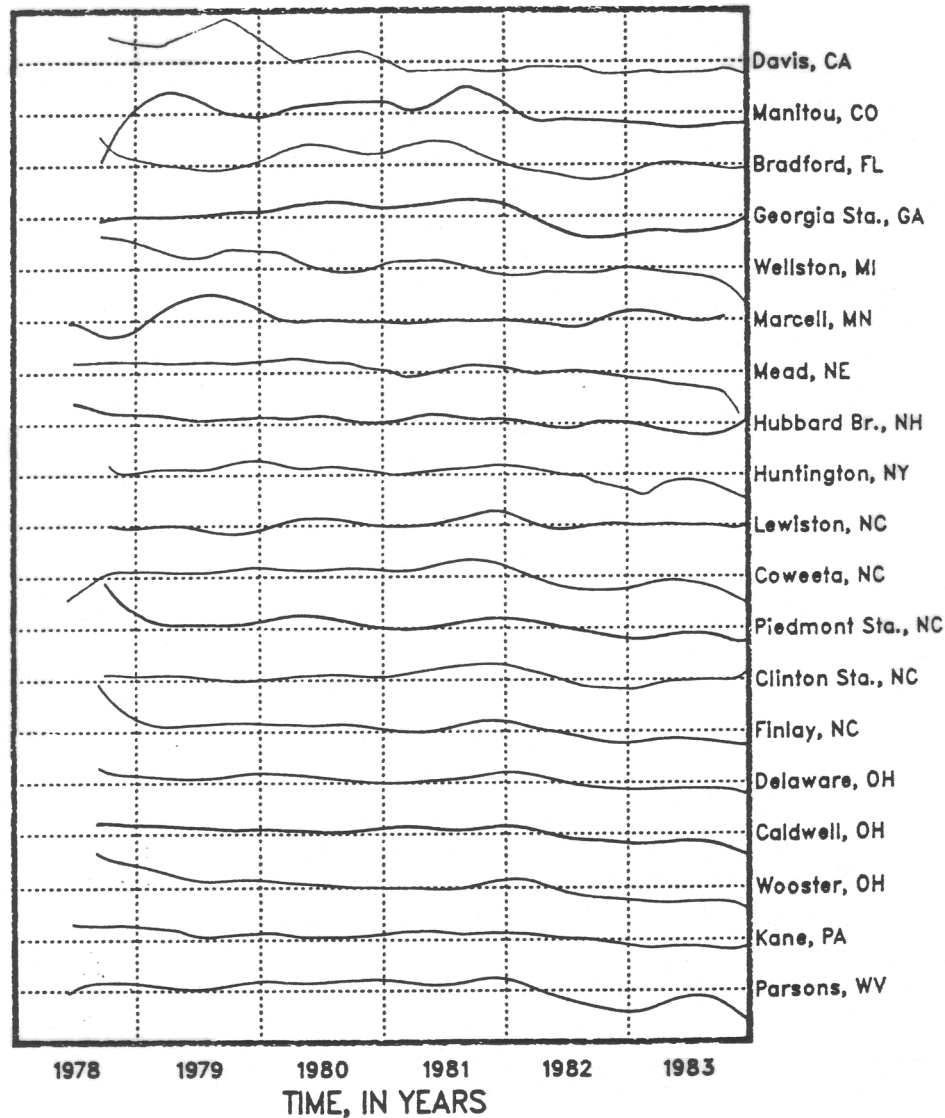


Figure 10.17 Smooths of sulfate concentrations at 19 stations, 1978-83
(from Schertz and Hirsch, 1985).

3. To remove the effect of an explanatory variable without first assuming the form of the relation (linear, etc.). In situations equivalent to multiple regression where several variables may affect the magnitude of a response variable (Y), removal of one variable's (X) effects may be accomplished by computing a LOWESS smooth of Y versus X and using the residuals from the smooth in subsequent analyses. An example is when removing the effects of discharge or precipitation volume from chemical concentration data prior to performing a trend analysis (see Chapter 12). LOWESS allows the analyst to be unconcerned as to whether the relation between Y and X is linear or nonlinear. In contrast, linearity would have to be established prior to using regression.

Two additional lines are sometimes plotted along with the LOWESS middle smooth. These are upper and lower smooths (Cleveland and McGill, 1984b), which function as smoothed versions of upper and lower quartiles of the conditional distribution of Y as a function of X . They are constructed by computing additional LOWESS smooths on the positive residuals and negative residuals, respectively, from the middle LOWESS smooth. These values are then added to the middle smooth, and connected with straight line segments. Upper and lower smooths are useful for showing how the spread and/or symmetry of the conditional distribution of Y changes as a function of X . Figure 10.18 is one example. It shows how the spread of nitrate concentrations changes with depth for groundwaters under Long Island, NY. The spread or "running IQR" is indicated by the distance between the upper and lower smooths, shown as dashed lines in the plot.

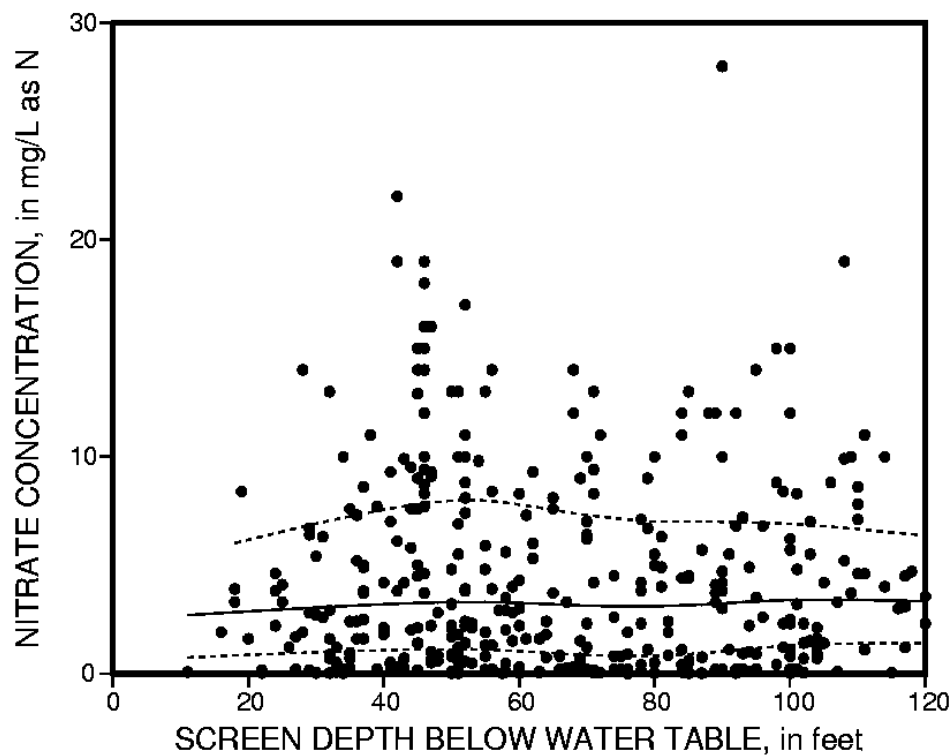


Figure 10.18 Nitrate concentrations versus depth in the upper Glacial Aquifer, Long Island NY (data from Eckhardt et al., 1989).

10.5.3 Polar Smoothing

Polar smooths (Cleveland and McGill, 1984b) are variations on lowess smooths. They are polygons describing the two-dimensional locations of data groups on a scatterplot (see figure 2.28 in Chapter 2). Comparisons of differences in location of several data groups is made much easier by comparing polar smooths rather than comparing symbols for each data point on a scatterplot, as in figure 2.27. Polar smooths are used as a visual 'discriminant analysis' in two dimensions.

To compute a polar smooth, first center the data at the median of X and median of Y. All data points are then described in terms of their angle and radius from this center, placing the data into polar coordinates. A lowess smooth is computed while in polar coordinates, and then is re-transformed back into original units. The smooth, which while in polar coordinates had 50 percent of the data below it, upon re-transformation envelops those same 50 percent within it. An analogous 'upper smooth' which in polar coordinates had 75 percent of the data below it becomes an 'outer smooth' containing 75 percent of the data in original units.

Polar smooths can be a great aid to exploratory data analysis. They are not constrained a priori to be an ellipse or any other shape, but take on the characteristics of the data. This can lead to new insights difficult to see by plotting the original observations. For example, in figure 2.28 smooths enclosing 75% of the conductance versus pH data for three types of upstream land use are plotted. The irregular pattern for the smooth of abandoned mine data suggests that two separate subgroups are present, one with higher pH than the other.

Exercises

- 10.1 For the data below,
- compute the Kendall slope estimator,
 - compute Kendall's τ ,
 - compute the non-parametric regression equation.
 - compute the significance level of the test.

Y	10	40	30	55	62	56
X	1	2	3	4	5	6

- 10.2 One value has been altered from the 10.1 exercise. Again compute the slope estimate, intercept, τ and significance level. By how much have these changed in response to the one (large) change in Y? Also compute a 95% confidence interval on the slope estimate.

Y	10	40	30	55	200	56
X	1	2	3	4	5	6

- 10.3 Compute the robust IWLS equation (2 iterations) for the Exercise 10.2 data.

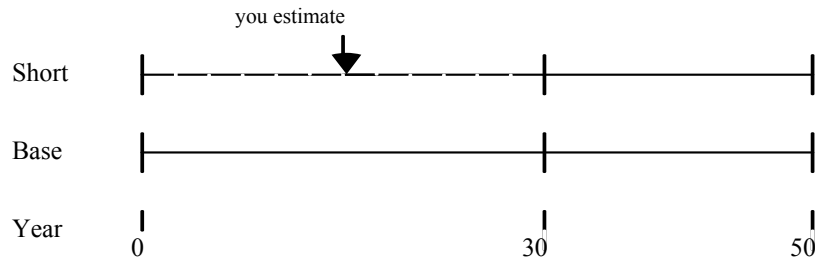
- 10.4 Williams and Wolman (1984) relate the lowering of streambed elevation downstream of a major dam to years following its installation. Calculate a linear least-squares regression of bed lowering (L) as the response variable, versus years (Yrs) as the explanatory variable, and compute its R^2 .

<u>Yrs</u>	<u>Lowering (m)</u>	<u>Yrs</u>	<u>L</u>	<u>Yrs</u>	<u>L</u>
0.5	-0.65	8	-4.85	17	-5.05
1	-1.20	10	-4.40	20	-5.10
2	-2.20	11	-4.95	22	-5.65
4	-2.60	13	-5.10	24	-5.50
6	-3.40	15	-4.90	27	-5.65

Calculate a 5-point moving median smooth of the data. Plot the smooth and regression line along with a scatterplot of the data. Describe how well each represents the data.

10.5 Record Extension

Monthly discharges for September at two rivers are given in Appendix C13 (units of million cubic meters per month). The most recent 20 years are available for "Short" (ignore the data in italics), and 50 years at "Base". The two sites are close enough that the data are reasonably well correlated with each other. Using the 20 years of joint record and the additional 30 years of record at "Base", produce a 50-year-long record at "Short" for use in a water supply simulation model.



First use regression and then repeat the process using the LOC. Take the extended record (the 30-year estimates plus the known 20 years) produced by the two methods at "Short" and plot them to illustrate the differences (a boxplot or probability plot are recommended). Compare these to each other and to a plot of the flows which actually occurred (the true flows are given in italics in Appendix C13). Which technique is preferable if the objective is to estimate water supply shortage risks? Which technique is preferable if the objective is to estimate the true September flow in each year? Quantify your conclusion about this.

- 10.6 The pulp liquor waste contamination of shallow groundwater (see Exercise 7.1) is revisited. Now the relationship between pH and COD in samples taken from the piezometers is of interest. Calculate a straight line which best describes the relationship between these two chemical constituents. Should this line be used by the field technician to predict COD from the pH measured on-site?

<u>pH</u>	<u>COD</u>	<u>pH</u>	<u>COD</u>	<u>pH</u>	<u>COD</u>
7.0	51	6.3	21	8.4	283
7.2	60	6.9	17	7.6	2170
7.5	51	7.0	34	7.5	6580
7.7	3600	6.4	43	7.4	3340
8.7	6900	6.8	34	9.3	7080
7.8	7700	6.7	43	9.0	10800